New Algorithms for M-Estimation of Multivariate Scatter and Location

Lutz Dümbgen[1*], Klaus Nordhausen[2**] and Heike Schuhmacher[1]
([1]University of Bern and [2]University of Turku)

May 2015, revised October 2015

Abstract

We present new algorithms for M-estimators of multivariate scatter and location and for symmetrized M-estimators of multivariate scatter. The new algorithms are considerably faster than currently used fixed-point and other algorithms. The main idea is to utilize a Taylor expansion of second order of the target functional and devise a partial Newton-Raphson procedure. In connection with symmetrized M-estimators we work with incomplete U-statistics to accelerate our procedures initially.

*Work supported by Swiss National Science Foundation.
**Work supported by Academy of Finland (grant 268703).

AMS subject classifications: 62H12, 65C60.

Key words: Fixed-point algorithm, matrix exponential function, Newton-Raphson algorithm, Taylor expansion.

Corresponding author: Lutz Dümbgen, e-mail: duembgen@stat.unibe.ch

1 Introduction

Robust estimation of multivariate location and scatter for a distribution P on $\mathbb{R}^q$ is a recurring topic in statistics. For instance, different estimators of multivariate scatter are an important ingredient for independent component analysis (ICA) or invariant coordinate selection (ICS), see Nordhausen et al. [10] and Tyler et al. [18] and the references therein. Of particular interest are M-estimators and their symmetrized versions as defined in Sections 2.1 and 2.3, respectively, because they offer a good compromise between robustness and computational feasibility. The most popular algorithm to compute M-estimators of multivariate scatter is to iterate a fixed-point equation, see Huber [7] (Section 8.11), Tyler [17] and Kent and Tyler [8]. This algorithm has nice properties such as guaranteed convergence for any starting point. However, as discussed later, it can be rather slow for high dimensions and large data sets. We introduce two alternative methods, a gradient descent method with approximately optimal stepsize and a partial Newton-Raphson method, which turn out to be substantially faster.

Computation time becomes a major issue in connection with symmetrized M-estimators. These estimators are important because of a desirable "block independence property" as explained in Section 2.3; see also Dümbgen [3] and Sirkiä et al. [16]. If applied to a sample of n observations $X_1, X_2, \ldots, X_n \in \mathbb{R}^q$, symmetrized M-estimators utilize the empirical distribution of all $\binom{n}{2}$ differences $X_i - X_j$, $1 \leq i < j \leq n$.

In Section 2 we describe briefly the various M-estimators we are interested in. Then we introduce a general target functional on the space of symmetric and positive definite matrices in $\mathbb{R}^{q\times q}$ which has to be minimized. Section 3 presents some analytical properties of the latter functional which are essential to understand existing algorithms and to devise new ones. These parts follow closely a recent survey of multivariate M-functionals by Dümbgen et al. [5]. In Section 4 we discuss the aforementioned fixed-point algorithm of Kent and Tyler [8] and explain rigorously why it is suboptimal. Then we introduce two alternative methods, a gradient descent method with approximately optimal stepsize and a partial Newton-Raphson method. Numerical experiments in Section 5 show that the new algorithms are substantially faster than the fixed-point algorithms or the algorithms by Arslan et al. [1]. Proofs are deferred to Section 6.

Some Notation. The space of symmetric matrices in $\mathbb{R}^{q\times q}$ is denoted by $\mathbb{R}^{q\times q}_{\mathrm{sym}}$, and $\mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ stands for its subset of positive definite matrices. The identity matrix in $\mathbb{R}^{q\times q}$ is written as I_q. The

Euclidean norm of a vector $v \in \mathbb{R}^q$ is denoted by $\|v\| = \sqrt{v^\top v}$. For matrices M, N with identical dimensions we write

$$\langle M, N \rangle \;:=\; \mathrm{tr}(M^\top N) \quad \text{and} \quad \|M\| \;:=\; \sqrt{\langle M, M \rangle},$$

so $\|M\|$ is the Frobenius norm of M.

2 The M-estimators and the target functional

Let $X_1, X_2, \ldots, X_n$ be independent random vectors with unknown distribution P on $\mathbb{R}^q$. Our task is to define and then estimate a certain center $\boldsymbol{\mu}(P) \in \mathbb{R}^q$ and scatter matrix $\boldsymbol{\Sigma}(P) \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$.

2.1 The scatter-only problem

Let us start with the assumption that $\boldsymbol{\mu}(P) = 0$. To define and estimate a scatter functional $\boldsymbol{\Sigma}(P)$ we consider a simple working model consisting of elliptically symmetric probability densities f_Σ on $\mathbb{R}^q$ depending on a parameter $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$:

$$f_\Sigma(x) \;=\; C^{-1} \det(\Sigma)^{-1/2} \exp\bigl(-\rho(x^\top \Sigma^{-1} x)/2\bigr),$$

where $\rho : [0,\infty) \to \mathbb{R}$ is a given function such that $C := \int \exp\bigl(-\rho(\|x\|^2)/2\bigr)\, dx$ is finite. Assuming temporarily that this working model is correct, one could estimate the true underlying matrix parameter by a maximizer of the corresponding log-likelihood function for this model,

$$\Sigma \;\mapsto\; -n \log C - \frac{1}{2} \sum_{i=1}^n \rho(X_i^\top \Sigma^{-1} X_i) - \frac{n}{2} \log\det(\Sigma).$$

With the empirical distribution $\widehat{P} = n^{-1} \sum_{i=1}^n \delta_{X_i}$ of the data $X_1, X_2, \ldots, X_n$, the log-likelihood at Σ may be written as $n \int \log f_\Sigma \, d\widehat{P}$. Thus maximization of the log-likelihood function over $\mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ is equivalent to minimization of $\Sigma \mapsto L(\Sigma, \widehat{P})$, where

$$\begin{aligned} L(\Sigma, Q) \;&:=\; 2 \int \log(f_{I_q} / f_\Sigma)\, dQ \\ &=\; \int \bigl[\rho(x^\top \Sigma^{-1} x) - \rho(x^\top x)\bigr]\, Q(dx) + \log\det(\Sigma) \end{aligned}$$

for a generic distribution Q on $\mathbb{R}^q$. We include f_{I_q} and $\rho(x^\top x)$, respectively, because often this increases the range of distributions Q such that $L(\Sigma, Q)$ is well-defined in $\mathbb{R}$. If $L(\cdot, Q)$ has a unique maximizer over $\mathbb{R}^{q\times q}_{\mathrm{sym},>0}$, we denote it with $\boldsymbol{\Sigma}(Q)$. The resulting mapping $Q \mapsto \boldsymbol{\Sigma}(Q)$ is called an M-functional of scatter. In particular, $\boldsymbol{\Sigma}(\widehat{P})$ serves as an estimator of the scatter

parameter $\mathbf{\Sigma}(P)$, assuming that both exist. If P happens to have a density f_{Σ_o} in our working model, then $\mathbf{\Sigma}(P) = \Sigma_o$. If P is merely elliptically symmetric with center 0 and scatter matrix Σ_o, for instance, if it has a density f of the form

$$f(x) \;=\; \det(\Sigma_o)^{-1/2} g_o(x^\top \Sigma_o^{-1} x)$$

with $g_o : [0,\infty) \to [0,\infty)$, then at least $\mathbf{\Sigma}(P) = \gamma\Sigma_o$ for some $\gamma > 0$.

An important example are multivariate t distributions with $\nu > 0$ degress of freedom. Here $\rho = \rho_{\nu,q}$ with

$$\rho_{\nu,q}(s) \;=\; (\nu+q)\log(\nu+s) \quad \text{for } s \geq 0. \tag{1}$$

Note that $\rho(x^\top\Sigma^{-1}x) - \rho(x^\top x)$ equals $(q+\nu)\log\big((\nu + x^\top\Sigma^{-1}x)/(\nu+x^\top x)\big)$, a bounded and smooth function of $x \in \mathbb{R}^q$.

2.2 The location-scatter problem

Now our working model consists of probability densities $f_{\mu,\Sigma}$ on $\mathbb{R}^q$ with parameters $\mu \in \mathbb{R}^q$ and $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$, namely,

$$f_{\mu,\Sigma}(x) \;=\; C^{-1}\det(\Sigma)^{-1/2}\exp\Big(-\rho\big((x-\mu)^\top\Sigma^{-1}(x-\mu)\big)/2\Big).$$

Here $(\boldsymbol{\mu}(P'), \mathbf{\Sigma}(P'))$ is defined as the minimizer of $2\int \log(f_{0,I_q}/f_{\mu,\Sigma})\,dP'$, where P' stands for P or $\widehat{P}$. But now we utilize a trick of Kent and Tyler [8] to get back to a scatter-only problem: With

$$y = y(x) \;:=\; \begin{bmatrix} x \\ 1 \end{bmatrix} \quad \text{and} \quad \Gamma = \Gamma(\mu,\Sigma) \;:=\; \begin{bmatrix} \Sigma + \mu\mu^\top & \mu \\ \mu^\top & 1 \end{bmatrix} \tag{2}$$

we may write $\log\det(\Sigma) = \log\det(\Gamma)$ and

$$-2\log f_{\mu,\Sigma}(x) \;=\; -2\log(C) + \rho(y^\top\Gamma^{-1}y - 1) + \log\det(\Gamma).$$

Hence $2\int \log(f_{0,I_q}/f_{\mu,\Sigma})\,dP'$ equals

$$L(\Gamma, Q) \;=\; \int \big[\rho(y^\top\Gamma^{-1}y - 1) - \rho(y^\top y - 1)\big]\,Q(dy) + \log\det(\Gamma)$$

with $Q := \mathcal{L}(y(X'))$, where $X' \sim P'$. Consequently, if $\mathbf{\Gamma} \in \mathbb{R}^{(q+1)\times(q+1)}_{\mathrm{sym},>0}$ minimizes $L(\cdot, Q)$ under the constraint

$$\mathbf{\Gamma}_{q+1,q+1} \;=\; 1,$$

then we may write

$$\mathbf{\Gamma} \;=\; \begin{bmatrix} \mathbf{\Sigma}(P') + \boldsymbol{\mu}(P')\boldsymbol{\mu}(P')^\top & \boldsymbol{\mu}(P') \\ \boldsymbol{\mu}(P')^\top & 1 \end{bmatrix},$$

and $(\boldsymbol{\mu}(P'), \boldsymbol{\Sigma}(P'))$ solves the original minimization problem. The mappings $P' \mapsto \boldsymbol{\mu}(P')$ and $P' \mapsto \boldsymbol{\Sigma}(P')$ are called M-functional of location and M-functional of scatter, respectively.

In the special case of $\rho = \rho_{\nu,q}$ with $\nu \geq 1$ we have the identity

$$\rho_{\nu,q}(s-1) \;=\; \rho_{\nu-1,q+1}(s) \quad \text{for } s > 0,$$

where we define

$$\rho_{0,q}(s) \;:=\; q\log(s) \quad \text{for } s > 0. \tag{3}$$

In case of $\nu > 1$ one can show that any minimizer $\boldsymbol{\Gamma}$ of $L(\cdot, Q)$ does satisfy the equation $\boldsymbol{\Gamma}_{q+1,q+1} = 1$, see [8] and [9]. In case of $\nu = 1$, which corresponds to multivariate Cauchy distributions, any minimizer $\boldsymbol{\Gamma}$ of $L(\cdot, Q)$ may be rescaled such that $\boldsymbol{\Gamma}_{q+1,q+1} = 1$. Thus in connection with multivariate t distributions with $\nu \geq 1$ degrees of freedom, the location-scatter problem can be reduced to a scatter-only problem.

If P has a density f_{μ_o,Σ_o} in our working model, then $(\boldsymbol{\mu}(P), \boldsymbol{\Sigma}(P)) = (\mu_o, \Sigma_o)$. If P is just elliptically symmetric with center μ_o and scatter matrix Σ_o, for instance, if it has a density f of the form

$$f(x) \;=\; \det(\Sigma_o)^{-1/2} g_o\big((x-\mu_o)^\top \Sigma_o^{-1}(x-\mu_o)\big)$$

with $g_o : [0,\infty) \to [0,\infty)$, then $\boldsymbol{\mu}(P) = \mu_o$ and $\boldsymbol{\Sigma}(P) = \gamma\Sigma_o$ for some $\gamma > 0$.

2.3 Symmetrized M-functionals

Suppose that P is (approximately) elliptically symmetric with unknown center μ_o and unknown scatter matrix Σ_o. In many situations one is only interested in the "shape matrix" $\det(\Sigma_o)^{-1/q}\Sigma_o$, i.e. a positive multiple of Σ_o with determinant 1. Examples are principal components, regression and correlation measures, where multiplying Σ_o with a positive scalar has no impact. Then we may get rid of the nuisance location parameter μ_o by replacing P with its symmetrization

$$P \ominus P \;:=\; \mathcal{L}(X' - X'') \quad \text{with independent } X', X'' \sim P.$$

Indeed, $P \ominus P$ is (approximately) elliptically symmetric with center 0 and the same shape matrix $\det(\Sigma_o)^{-1/q}\Sigma_o$. We may estimate $P \ominus P$ by the measure-valued U-statistic

$$\widehat{P \ominus P} \;:=\; \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \delta_{X_i - X_j}.$$

Then, if we define $\mathbf{\Sigma}(Q)$ to be the minimizer of

$$\int \left[\rho(x^\top \Sigma^{-1} x) - \rho(x^\top x)\right] Q(dx) + \log\det(\Sigma)$$

with respect to Σ, then the shape matrix of $\mathbf{\Sigma}(\widehat{P \ominus P})$ is a plausible estimator of the true shape matrix $\det(\Sigma_o)^{-1/q}\Sigma_o$. The mapping $P \mapsto \mathbf{\Sigma}(P \ominus P)$ is called a symmetrized M-functional of scatter.

This symmetrization has a second, even more important advantage: Consider an arbitrary distribution P, i.e. it may fail to be (approximately) elliptically symmetric. But suppose that a random vector $X \sim P$ may be written as $X = [X_1^\top, X_2^\top]^\top$ with independent subvectors $X_1 \in \mathbb{R}^{q(1)}, X_2 \in \mathbb{R}^{q(2)}$. Then $\mathbf{\Sigma}(P)$ is block-diagonal in the sense that

$$\mathbf{\Sigma}(P) \;=\; \begin{bmatrix} \mathbf{\Sigma}_1(P) & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2(P) \end{bmatrix}$$

with symmetric matrices $\mathbf{\Sigma}_i(P) \in \mathbb{R}_{\mathrm{sym}}^{q(i)\times q(i)}$. For a further discussion on the use of symmetrized scatter matrices in multivariate statistics see also Nordhausen and Tyler [13].

2.4 The general settings

Let Q be a probability distribution on $\mathbb{R}^q$. Now we seek to minimize a certain target functional $L(\cdot, Q)$ on the space $\mathbb{R}_{\mathrm{sym},>0}^{q\times q}$ of symmetric and positive definite matrices in $\mathbb{R}^{q\times q}$, where $L(\cdot,\cdot)$ and Q have to satisfy certain conditions:

Setting 0. We assume that $Q(\{0\}) = 0$, and for $\Sigma \in \mathbb{R}_{\mathrm{sym},>0}^{q\times q}$ we define

$$L_0(\Sigma, Q) \;:=\; q \int \log\Big(\frac{x^\top \Sigma^{-1} x}{x^\top x}\Big)\, Q(dx) + \log\det(\Sigma).$$

Moreover, we assume that

$$Q(\mathbb{V}) \;<\; \frac{\dim(\mathbb{V})}{q}$$

for any linear subspace $\mathbb{V}$ of $\mathbb{R}^q$ with $1 \le \dim(\mathbb{V}) < q$.

Setting 1. Let $\rho : [0,\infty) \to \mathbb{R}$ be twice continuously differentiable such that $\rho' > 0 \ge \rho''$. Further we assume that $\psi(s) := s\rho'(s)$ satisfies the following two properties: $\psi' > 0$ and $q < \psi(\infty) := \lim_{s\to\infty} \psi(s) < \infty$. For $\Sigma \in \mathbb{R}_{\mathrm{sym},>0}^{q\times q}$ we define

$$L_\rho(\Sigma, Q) \;:=\; \int \left[\rho(x^\top \Sigma^{-1} x) - \rho(x^\top x)\right] Q(dx) + \log\det(\Sigma).$$

Moreover, we assume that

$$Q(\mathbb{V}) \;<\; \frac{\psi(\infty) - q + \dim(\mathbb{V})}{\psi(\infty)}$$

for any linear subspace $\mathbb{V}$ of $\mathbb{R}^q$ with $0 \leq \dim(\mathbb{V}) < q$.

Note that for $\nu > 0$, $\rho = \rho_{\nu,q}$ satisfies the conditions of Setting 1 with $\psi(s) = (\nu+q)s/(\nu+s)$. Hence $\psi(\infty) = \nu + q$, and Q has to satisfy

$$Q(\mathbb{V}) \ < \ \frac{\nu + \dim(\mathbb{V})}{\nu + q}$$

for proper linear subspaces $\mathbb{V}$ of $\mathbb{R}^q$.

Note also that Setting 0 is similar to Setting 1 if we define $\rho := \rho_{0,q}$ as in (3). The main difference to Setting 1 is that $L_0(t\Sigma, Q) = L_0(\Sigma, Q)$ for arbitrary $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ and $t > 0$. In what follows we often write $L(\Sigma, Q)$ for $L_0(\Sigma, Q)$ or $L_\rho(\Sigma, Q)$.

The assumptions on ρ and Q imply that the functional $L(\cdot, Q)$ has essentially a unique minimizer (see [8], [2] or [5]):

Theorem 1. *In Setting 0 there exists a unique matrix $\mathbf{\Sigma}_0(Q) \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ such that*

$$L_0(\mathbf{\Sigma}_0(Q), Q) \leq L_0(\cdot, Q) \quad \textit{and} \quad \det(\mathbf{\Sigma}_0(Q)) \ = \ 1.$$

In Setting 1 there exists a unique matrix $\mathbf{\Sigma}_\rho(Q) \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ such that

$$L_\rho(\mathbf{\Sigma}_\rho(Q), Q) \ \leq \ L_\rho(\cdot, Q).$$

Coming back to the specific situation with independent random variables $X_1, X_2, \ldots, X_n$ with distribution P on $\mathbb{R}^q$, the scatter estimators in Sections 2.1, 2.2 and 2.3 correspond to the following choices of Q:

- $Q = \widehat{P} = n^{-1} \sum_{i=1}^n \delta_{X_i}$ (Section 2.1);
- $Q = n^{-1} \sum_{i=1}^n \delta_{y(X_i)}$ with dimension $q+1$ in place of q (Section 2.2);
- $Q = \widehat{P \ominus P} = \binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} \delta_{X_i - X_j}$ (Section 2.3).

3 Analytical properties of $L(\cdot, Q)$

As shown in Dümbgen et al. [5], the functionals $L_0(\cdot, Q)$ and $L_\rho(\cdot, Q)$ are smooth, strictly convex and coercive in a certain sense. To make this precise, we utilize the matrix-valued exponential function: For $A \in \mathbb{R}^{q\times q}$ let

$$\exp(A) \ := \ \sum_{k=0}^\infty \frac{1}{k!} A^k.$$

In case of $A = A^\top$ we may write $A = U \operatorname{diag}(\lambda) U^\top$ with an orthogonal matrix $U \in \mathbb{R}^{q\times q}$ and some vector $\lambda = (\lambda_i)_{i=1}^q \in \mathbb{R}^q$. Then

$$\exp(A) \;=\; U \operatorname{diag}(\exp(\lambda)) U^\top$$

with $\exp(\lambda) := \big(\exp(\lambda_i)\big)_{i=1}^q$. Moreover,

$$\log\det(\exp(A)) \;=\; \operatorname{tr}(A).$$

If $A \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$, i.e. $\lambda \in (0,\infty)^q$, then $A = \exp(\log(A))$ with

$$\log(A) \;:=\; U \operatorname{diag}(\log(\lambda)) U^\top$$

and $\log(\lambda) := (\log \lambda_i)_{i=1}^q$.

By means of the matrix-valued exponential function and logarithm, we can describe the behavior of $L(\cdot, Q)$ in a neighborhood of any matrix $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ quite elegantly. Instead of considering additive perturbations $\Sigma + A$ with $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$, we write $\Sigma = BB^\top$ for some nonsingular matrix $B \in \mathbb{R}^{q\times q}$, for instance $B = \Sigma^{1/2}$, and consider multiplicative perturbations $B\exp(A)B^\top$. Note that

$$\big\{B\exp(A)B^\top : A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}\big\} \;=\; \mathbb{R}^{q\times q}_{\mathrm{sym},>0}.$$

In case of $\det(\Sigma) = 1$,

$$\big\{B\exp(A)B^\top : A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}, \operatorname{tr}(A) = 0\big\} \;=\; \big\{\Gamma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0} : \det(\Gamma) = 1\big\}.$$

Here is a basic expansion of $L\big(B\exp(\cdot)B^\top, Q\big)$ around 0:

Theorem 2 ([5]). *For a nonsingular matrix $B \in \mathbb{R}^{q\times q}$ define $Q_B := \mathcal{L}(B^{-1}X)$ with $X \sim Q$. Then for $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$,*

$$\begin{aligned} &L\big(B\exp(A)B^\top, Q\big) - L\big(BB^\top, Q\big) \\ &\quad=\; L(\exp(A), Q_B) \;=\; G(A, Q_B) + \frac{1}{2} H(A, Q_B) + o(\|A\|^2) \end{aligned}$$

as $A \to 0$, where

$$\begin{aligned} G(A, Q_B) \;&:=\; \big\langle A, I_q - \Psi(Q_B)\big\rangle, \\ H(A, Q_B) \;&:=\; \big\langle A^2, \Psi(Q_B)\big\rangle + \int \rho''(\|x\|^2)(x^\top A x)^2 \, Q_B(dx), \end{aligned}$$

and

$$\Psi(Q_B) \;:=\; \int \rho'(\|x\|^2)\, xx^\top \, Q_B(dx).$$

Moreover, $H(A, Q_B)$ is continuous in B, and

$$H(A, Q_B) \begin{cases} \geq 0, & \\ > 0 & \textit{in Setting 0, if } A \notin \{sI_q : s \in \mathbb{R}\}, \\ > 0 & \textit{in Setting 1, if } A \neq 0. \end{cases}$$

Remark 3. The Taylor expansion in Theorem 2 implies that

$$L(B\exp(A)B^\top, Q) \;=\; L(B\exp(0)B^\top, Q) + \langle A, G(Q_B)\rangle + O(\|A\|^2)$$

as $A \to 0$, where

$$G(Q_B) \;:=\; I_q - \Psi(Q_B) \;\in\; \mathbb{R}^{q\times q}_{\mathrm{sym}}.$$

Hence the matrix $G(Q_B)$ is the gradient of the function $\mathbb{R}^{q\times q}_{\mathrm{sym}} \ni A \mapsto L\big(B\exp(\cdot)B^\top, Q\big)$ at $0 \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$.

Note also that $\Psi(Q_B)$ is positive definite, because otherwise Q would be concentrated on a proper linear subspace of $\mathbb{R}^q$.

Remark 4. Note that $L_0(t\Sigma, Q)$ is constant in $t > 0$ for any $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$. In other words, for any nonsingular $B \in \mathbb{R}^{q\times q}$, $L_0(B\exp(xI_q)B^\top, Q)$ is constant in $x \in \mathbb{R}$. Applying Theorem 2 to $A = xI_q$ yields that $G(I_q, Q_B) = \mathrm{tr}(G(Q_B)) = 0$ and $H(I_q, Q_B) = 0$ in Setting 0. This explains the constraint $A \notin \{sI_q : s \in \mathbb{R}\}$ for $H(A, Q_B) > 0$.

Remark 5. The second derivative of the function $L\big(B\exp(\cdot)B^\top, Q\big)$ at $0 \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$ corresponds to the quadratic form

$$\mathbb{R}^{q\times q}_{\mathrm{sym}} \times \mathbb{R}^{q\times q}_{\mathrm{sym}} \ni (A', A) \;\mapsto\; \big\langle A', H(Q_B)A\big\rangle$$

with the self-adjoint linear operator $H(Q_B) : \mathbb{R}^{q\times q}_{\mathrm{sym}} \to \mathbb{R}^{q\times q}_{\mathrm{sym}}$ given by

$$H(Q_B)A \;:=\; 2^{-1}\big(\Psi(Q_B)A + A\Psi(Q_B)\big) + \int \rho''(\|x\|^2)x^\top Ax\, xx^\top\, Q_B(dx).$$

Theorem 2 implies that this operator is positive definite in Setting 1. In Setting 0,

$$\begin{aligned} \Psi(Q_B) \;&=\; q\int \|x\|^{-2}\, xx^\top\, Q_B(dx), \\ H(Q_B)A \;&=\; 2^{-1}\big(\Psi(Q_B)A + A\Psi(Q_B)\big) - q\int \|x\|^{-4}x^\top Ax\, xx^\top\, Q_B(dx), \end{aligned}$$

and one easily verifies that $H(Q_B)I_q = 0$ and $\mathrm{tr}\big(H(Q_B)A\big) = 0$ for any $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$. Hence in both settings one may view $H(Q_B)$ as a self-adjoint and positive definite linear operator from the set

$$\mathbb{W} \;:=\; \begin{cases} \big\{A \in \mathbb{R}^{q\times q}_{\mathrm{sym}} : \mathrm{tr}(A) = 0\big\} & \text{in Setting 0} \\ \mathbb{R}^{q\times q}_{\mathrm{sym}} & \text{in Setting 1} \end{cases}$$

onto itself. In particular, $H(Q_B)^{-1}$ stands for the corresponding inverse mapping.

An important consequence of Theorem 2 is a convexity property of $L(\cdot, Q)$:

Corollary 6. *For any nonsingular $B \in \mathbb{R}^{q\times q}$ and $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$, the mapping*

$$t \;\mapsto\; L\big(B\exp(tA)B^\top, Q\big)$$

is twice continuously differentiable and convex on $\mathbb{R}$. In Setting 0 it is strictly convex if $A \notin \{sI_q : s \in \mathbb{R}\}$. In Setting 1 it is strictly convex if $A \neq 0$.

This corollary implies that $\Sigma = BB^\top$ minimizes $L(\cdot, Q)$ if, and only if, the gradient $G(Q_B)$ equals 0, i.e.

$$\Psi(Q_B) \;=\; I_q. \tag{4}$$

This is equivalent to the fixed-point equation

$$\Sigma \;=\; \int \rho'(x^\top\Sigma^{-1}x)xx^\top\, Q(dx). \tag{5}$$

4 Algorithms

4.1 Fixed-point and gradient algorithms

The fixed-point equation (5) gives rise to a fixed-point algorithm which has been proposed and used repeatedly, see for instance Huber [7] (Section 8.11), Tyler [17] and Kent and Tyler [8]. The latter two references provide a rigorous proof of convergence for empirical distributions Q, the general case is covered by Dudley et al. [2]. A basic step works as follows: If $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ is our current candidate for a minimizer of $L(\cdot, Q)$, then we replace it with

$$\int \rho'(x^\top\Sigma^{-1}x)xx^\top\, Q(dx).$$

When implementing this method it is more convenient to utilize the formulation (4) directly: If $\Sigma = BB^\top$ for some nonsingular matrix $B \in \mathbb{R}^{q\times q}$, then

$$\int \rho'(x^\top\Sigma^{-1}x)xx^\top\, Q(dx) \;=\; B\Psi(Q_B)B^\top.$$

Now we use some factorization $\Psi(Q_B) = CC^\top$ with nonsingular $C \in \mathbb{R}^{q\times q}$ and replace B with BC. Replacing Σ with $B\Psi(Q_B)B^\top$ yields always an improvement, because

$$L(B\Psi(Q_B)B^\top, Q) - L(BB^\top, Q) \;<\; 0 \quad \text{unless } \Psi(Q_B) = I_q; \tag{6}$$

see [5]. Here is a description of the fixed-point algorithm:

Algorithm FP. Choose an arbitrary matrix $\Sigma_0 = B_0 B_0^\top$ with nonsingular $B_0 \in \mathbb{R}^{q\times q}$, and let $Q_0 := Q_{B_0}$. Suppose that after $k \geq 0$ steps we have determined a nonsingular matrix $B_k \in \mathbb{R}^{q\times q}$, corresponding to the candidate $\Sigma_k = B_k B_k^\top$ for $\mathbf{\Sigma}(Q)$. Writing $Q_k := Q_{B_k}$, we compute

$$\Psi_k \;:=\; \Psi(Q_k) \;=\; \int \rho'(\|x\|^2)\, xx^\top\, Q_k(dx).$$

Then we write $\Psi_k = C_k C_k^\top$ for some nonsingular $C_k \in \mathbb{R}^{q\times q}$ and define

$$B_{k+1} \;:=\; B_k C_k.$$

This corresponds to the new candidate $\Sigma_{k+1} := B_{k+1} B_{k+1}^\top = B_k \Psi_k B_k^\top$.

This description is similar to the one of Huber [7] (Section 8.11), the main difference being that we don't restrict ourselves to the Cholesky factorization of Ψ_k. Indeed in our implementation we use $\Psi_k = C_k C_k^\top$ with $C_k = U_k \operatorname{diag}(\phi_k)^{1/2}$, where $\phi_k \in (0,\infty)^q$ contains the eigenvalues of Ψ_k and U_k is an orthogonal matrix of corresponding eigenvectors. Our starting point is typically

$$\Sigma_0 \;:=\; \int xx^\top\, Q(dx).$$

Our stopping criterion for Algorithm FP is that $\|I_q - \Psi_k\| = \|1_q - \phi_k\| < \delta$ for some given small number $\delta > 0$, where $1_q := (1, 1, \ldots, 1)^\top \in \mathbb{R}^q$.

An important fact is that under the conditions of Theorem 1 the sequence $(\Sigma_k)_{k=0}^\infty$ converges to a minimizer of $L(\cdot, Q)$, no matter which starting point Σ_0 has been chosen; see also Theorem 8 later.

One may view the fixed-point algorithm as an approximate gradient method with constant stepsize one: Note that with the gradient $G_k := G(Q_k)$ of $L(B_k \exp(\cdot) B_k^\top, Q)$ at $0 \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$,

$$\Sigma_{k+1} \;=\; B_k \Psi_k B_k^\top \;=\; B_k (I_q - G_k) B_k^\top \;=\; B_k \exp\bigl(-G_k + O(\|G_k\|^2)\bigr) B_k^\top.$$

In the present context an exact gradient method with constant step size one would mean to replace Σ_k with $B_k \exp(-G_k) B_k^\top$.

Suboptimality of Algorithm FP. As shown later, the steps performed in Algorithm FP are clearly suboptimal, at least when Σ_k is already close to the limit $\mathbf{\Sigma}(Q)$. To understand this thoroughly and to devise improvements we first provide a corollary to Theorem 2:

Corollary 7. *Let $\Sigma = BB^\top$ for a nonsingular matrix $B \in \mathbb{R}^{q\times q}$. Further let $Q_* := Q_{\mathbf{\Sigma}(Q)^{1/2}}$. If we write $B = \Sigma^{1/2}V$ with an orthogonal matrix $V \in \mathbb{R}^{q\times q}$, then for any $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$,*

$$\begin{aligned} L(B\exp(A)B^\top, Q) - L(BB^\top, Q) &= L(\exp(A), Q_B) \\ &= G(A, Q_B) + \frac{1}{2}H(A, Q_B) + r(B, A)\|A\|^2 \\ &= G(A, Q_B) + \frac{1}{2}H(V^\top AV, Q_*) + r_*(B, A)\|A\|^2, \end{aligned}$$

where

$$|r(B, A)| + |r_*(B, A)| \;\to\; 0 \quad \text{as } BB^\top \to \mathbf{\Sigma}(Q) \text{ and } A \to 0.$$

Moreover,

$$H(V^\top AV, Q_*) \;=\; \|A\|^2 + \int \rho''(\|x\|^2)(x^\top V^\top AVx)^2\, Q_*(dx).$$

Now let us apply this corollary to Algorithm FP. We write $B_k = \Sigma_k^{1/2}V_k$ for some orthogonal matrix $V_k \in \mathbb{R}^{q\times q}$. If we fix an arbitrary constant $K > 1$, then uniformly in $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$ with $\|A\| \le K\|G_k\|$,

$$\begin{aligned} L(B_k\exp(A)B_k^\top, Q) - L(B_kB_k^\top, Q) &= L(\exp(A), Q_k) \\ &= \langle A, G_k\rangle + \frac{1}{2}H(V_k^\top AV_k, Q_*) + r_*(B_k, A)\|A\|^2 \\ &= \langle A, G_k\rangle + \frac{1}{2}H(V_k^\top AV_k, Q_*) + o(\|G_k\|^2). \end{aligned}$$

In particular, if we choose $A = -t_kG_k$ with a bounded sequence $(t_k)_k$ in $\mathbb{R}$,

$$L(\exp(-t_kG_k), Q_k) \;=\; \|G_k\|^2\Big(-t_k + \frac{t_k^2}{2}\,\frac{H(V_k^\top G_kV_k, Q_*)}{\|G_k\|^2} + o(1)\Big).$$

Consequently, an approximately optimal choice of t_k would be a minimizer of the right hand side without the term $o(1)$, i.e.

$$\begin{aligned} t_k^* &= \frac{\|G_k\|^2}{H(V_k^\top G_kV_k, Q_*)} \\ &= \Big(1 + \int \rho''(\|x\|^2)\frac{(x^\top V_k^\top G_kV_kx)^2}{\|G_k\|^2}\, Q_*(dx)\Big)^{-1} \\ &\in \Big[\Big(1 - \min_{A\in\mathbb{W}:\|A\|=1}\int |\rho''|(\|x\|^2)(x^\top Ax)^2\, Q_*(dx)\Big)^{-1}, \lambda_{\min}\big(H(Q_*)\big)^{-1}\Big]. \end{aligned}$$

The upper bound involves the minimal eigenvalue of the symmetric operator $H(Q_*) : \mathbb{W} \to \mathbb{W}$. The lower bound follows from $\rho'' \le 0$ and is typically strictly larger than 1, for instance if $\rho = \rho_{\nu,q}$ as defined in (1) or (3). Hence the steps performed during the fixed-point algorithm tend to be too short!

Algorithm G. One could easily fix this deficiency as follows: As a proxy for t_k^*, which involves the unknown quadratic form $H(\cdot, Q_*)$, we compute in the k-th iteration the number

$$t_k \;=\; \frac{\|G_k\|^2}{H(G_k, Q_k)} \;=\; t_k^*\,(1 + o(1)).$$

The latter equality follows from Corollary 7. Indeed, the latter corollary implies that we obtain $L(\exp(-t_k G_k), Q_k) \;=\; -\|G_k\|^4/(2H(G_k, Q_k))(1 + o(1)) \;\leq\; -\|G_k\|^2/2(1 + o(1))$. Thus we check whether

$$L\big(\exp(-t_k G_k), Q_k\big) \;\leq\; -\|G_k\|^2/4. \tag{7}$$

If yes, we replace B_k with $B_{k+1} = B_k C_k$, where $C_k C_k^\top = \exp(-t_k G_k)$. Otherwise we perform a usual fixed-point step as described before. The number 4 in (7) could be replaced with any number $c > 2$.

Implementing this gradient method yielded already a substantial reduction of computation time. This approach of improving a fixed-point algorithm by means of variable step lengths is also used by Redner and Walker [15] in the context of maximum-likelihood estimation for mixture models. But in view of Theorem 2 it is certainly tempting to try a Newton-Raphson procedure.

4.2 (Partial) Newton-Raphson procedures

Suppose that our current candidate for $\mathbf{\Sigma}(Q)$ is $\Sigma \;=\; BB^\top$. In view of Corollary 7 we should replace Σ with

$$\tilde{\Sigma} \;=\; B \exp\big(-H(Q_B)^{-1} G(Q_B)\big) B^\top,$$

because $H(Q_B)^{-1} G(Q_B)$ is the unique minimizer of

$$\mathbb{W} \ni A \;\mapsto\; G(A, Q_B) + \frac{1}{2} H(A, Q_B) \;=\; \langle A, G(Q_B)\rangle + \frac{1}{2}\langle A, H(Q_B)A\rangle.$$

A problem with this promising update $\tilde{\Sigma}$ is that the computation of the inverse operator $H(Q_B)^{-1}$ may be too computer- or memory-intensive. Indeed, we implemented a full Newton-Raphson algorithm, and it required only very few iterations, as expected. But the running time was even longer than with Algorithm FP, because the computation and inversion of $H(Q_B)$, which may be represented by a symmetric matrix in $\mathbb{R}^{\dim(\mathbb{W})\times\dim(\mathbb{W})}$, was too time-consuming. Note that $\dim(\mathbb{W})$ equals $q(q+1)/2 - 1$ in Setting 0 and $q(q+1)/2$ in Setting 1.

These difficulties with a full Newton-Raphson procedure have been noticed already by Huber [7] (Section 8.11). Some authors have tried alternative approaches such as conjugate gradient

methods or quasi Newton methods in which the operator $H(Q_B)$ is replaced with a surrogate which is easier to compute and invert; see for instance Huber [6]. According to [7], none of these attempts was overall convincing.

A partial Newton-Raphson approach turned out to be quite successful. This means that instead of considering arbitrary multiplicative perturbations $B\exp(A)B^\top$ of a current candidate $\Sigma = BB^\top$, we restrict A to a particular q-dimensional subspace of $\mathbb{R}^{q\times q}_{\mathrm{sym}}$ depending on B. Precisely, consider the matrix $\Psi(Q_B) \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ and its spectral decomposition,

$$\Psi(Q_B) \;=\; U\,\mathrm{diag}(\phi)U^\top$$

with an orthogonal matrix $U \in \mathbb{R}^{q\times q}$ whose columns are eigenvectors of $\Psi(Q_B)$ and a vector $\phi \in (0,\infty)^q$ containing the corresponding eigenvalues. Now we consider only perturbations $\Sigma = B\exp(A)B^\top$ with $A = U\,\mathrm{diag}(a)U^\top$, $a \in \mathbb{R}^q$. Since $\exp(U\,\mathrm{diag}(a)U^\top) = U\exp(\mathrm{diag}(a))U^\top$, this leads to the functional

$$\mathbb{R}^q \ni a \;\mapsto\; L\big(B\,U\exp(\mathrm{diag}(a))U^\top B^\top, Q\big) - L(BB^\top, Q).$$

Now the Taylor expansion in Theorem 2 may be rewritten as follows:

$$\begin{aligned} &L\big(B\,U\exp(\mathrm{diag}(a))U^\top B^\top, Q\big) - L(BB^\top, Q) \\ &\qquad =\; L\big(\exp(\mathrm{diag}(a)), Q_{BU}\big) \;=\; \tilde G(Q_{BU})^\top a + \frac{1}{2}\,a^\top \tilde H(Q_{BU})a + o(\|a\|^2), \end{aligned}$$

where

$$\begin{aligned} \tilde G(Q_{BU}) \;&:=\; 1_q - \int \rho'(\|x\|^2)s(x)\,Q_{BU}(dx) \;=\; 1_q - \phi \;\in\; \mathbb{R}^q, \\ \tilde H(Q_{BU}) \;&:=\; \mathrm{diag}(\phi) + \int \rho''(\|x\|^2)s(x)s(x)^\top\,Q_{BU}(dx) \;\in\; \mathbb{R}^{q\times q}_{\mathrm{sym}} \end{aligned}$$

with $1_q = (1)_{j=1}^q$ and

$$s(x) \;:=\; (x_j^2)_{j=1}^q \quad \text{for } x = (x_j)_{j=1}^q \in \mathbb{R}^q.$$

In Setting 1, $\tilde H(Q_{BU})$ is a positive definite matrix, and

$$\operatorname*{arg\,min}_{a\in\mathbb{R}^q}\big(\tilde G(Q_{BU})^\top a + \frac{1}{2}\,a^\top \tilde H(Q_{BU})a\big) \;=\; -\tilde H(Q_{BU})^{-1}\tilde G(Q_{BU}).$$

In Setting 0, the matrix $\tilde H(Q_{BU})$ satisfies $\tilde H(Q_{BU})1_q \;=\; 0$ and $a^\top \tilde H(Q_{BU})a \;>\; 0$ whenever $a \neq 0$, $1_q^\top a = 0$. Moreover, $1_q^\top \tilde G(Q_{BU}) = 0$. Thus we may write

$$\operatorname*{arg\,min}_{a\in\mathbb{R}^q}\big(\tilde G(Q_{BU})^\top a + \frac{1}{2}\,a^\top \tilde H(Q_{BU})a\big) \;=\; -(\tilde H(Q_{BU}) + c\,1_q 1_q^\top)^{-1}\tilde G(Q_{BU})$$

for any constant $c > 0$.

Algorithm PN. Choose an arbitrary matrix $\Sigma_0 = B_0 B_0^\top$ with nonsingular $B_0 \in \mathbb{R}^{q\times q}$, and let $Q_0 := Q_{B_0}$.

Suppose that for some integer $k \geq 0$ we have already determined a nonsingular matrix $B_k \in \mathbb{R}^{q\times q}$. Writing $Q_k := Q_{B_k}$, we compute

$$\Psi_k \;:=\; \Psi(Q_k) \;=\; \int \rho'(\|x\|^2)\, xx^\top\, Q_k(dx).$$

Then we write $\Psi_k = U_k \operatorname{diag}(\phi_k) U_k^\top$ with an orthogonal matrix $U_k \in \mathbb{R}^{q\times q}$ and a vector $\phi_k \in (0,\infty)^q$. Next we define

$$\tilde{Q}_k \;:=\; (Q_k)_{U_k} \;=\; Q_{B_k U_k}$$

and

$$a_k \;:=\; \begin{cases} -\tilde{H}(\tilde{Q}_k)^{-1}\tilde{G}(\tilde{Q}_k) & \text{in Setting 1,} \\ -\big(\tilde{H}(\tilde{Q}_k) + c\, 1_q 1_q^\top\big)^{-1}\tilde{G}(\tilde{Q}_k) & \text{in Setting 0.} \end{cases}$$

We expect that replacing B_k with $B_k \exp(\operatorname{diag}(a_k/2))$ results in a change of $L(\cdot, Q)$ of about $a_k^\top \tilde{G}(\tilde{Q}_k)/2 < 0$. Now we check whether

$$L\big(\exp(\operatorname{diag}(a_k)), \tilde{Q}_k\big) \;\leq\; a_k^\top \tilde{G}(\tilde{Q}_k)/4. \tag{8}$$

If yes, we define

$$B_{k+1} \;:=\; B_k U_k \exp(\operatorname{diag}(a_k/2))$$

which corresponds to the new candidate $\Sigma_{k+1} := B_{k+1}B_{k+1}^\top = B_k \exp(\operatorname{diag}(a_k)) B_k^\top$. If (8) is violated we just perform a step of the fixed-point algorithm and set $B_{k+1} := B_k U_k \operatorname{diag}(\phi_k)^{1/2}$, i.e. our new candidate is $\Sigma_{k+1} := B_{k+1}B_{k+1}^\top = B_k \operatorname{diag}(\phi_k) B_k^\top$. Again, the number 4 in (8) could be replaced by any number $c > 2$.

The new Algorithm PN is also guaranteed to converge to a minimizer of $L(\cdot, Q)$:

Theorem 8. *For any starting point $\Sigma_0 \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$ and in both Settings 0 and 1, Algorithm FP as well as Algorithm PN yield a sequence $(\Sigma_k)_k$ converging to a minimizer Σ_* of $L(\cdot, Q)$.*

For general distributions Q it is difficult to compare Algorithms FP and PN explicitly. Recall that in Algorithm PN we restrict our attention to a particular subspace of $\mathbb{R}^{q\times q}_{\mathrm{sym},>0}$. The following lemma implies that at least in case of an (approximately) elliptically symmetric distribution Q this subspace is (almost) the right one to look in for better candidates.

Lemma 9. *Suppose that Q is elliptically symmetric with center 0 and scatter matrix $\Sigma_o \in \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$. Then $\boldsymbol{\Sigma}(Q) = \kappa\Sigma_o$ for some $\kappa > 0$. Moreover, for any $\Sigma = BB^\top$ with nonsingu-*

lar $B \in \mathbb{R}^{q\times q}$ and any spectral decomposition $\Psi(Q_B) = U\,\mathrm{diag}(\phi)U^\top$,

$$\boldsymbol{\Sigma}(Q) \;=\; BU\exp(\mathrm{diag}(a))U^\top B^\top$$

for a vector $a \in \mathbb{R}^q$ containing the log-eigenvalues of $\Sigma^{-1}\boldsymbol{\Sigma}(Q)$.

At this point we should mention that for "well-behaved" distributions Q in high dimension q, algorithm FP can be rather efficient, because the standardized distribution $Q_* = Q_{\boldsymbol{\Sigma}(Q)^{1/2}}$ satisfies

$$H(A, Q_*) \;\approx\; \|A\|^2$$

for $A \in \mathbb{W}$. For instance in Setting 0, if Q_* is spherically symmetric around 0,

$$H(A, Q_*) \;=\; \frac{q}{q+2}\|A\|^2$$

for all $A \in \mathbb{W}$. Hence, if $\Sigma = BB^\top$ is already close to $\boldsymbol{\Sigma}(Q)$, the Newton step would be to replace Σ with

$$\Sigma_{\mathrm{new}} \;\approx\; B\exp(-(1+2/q)G(Q_B))B^\top,$$

and for high dimension q this is similar to $B\exp(-G(Q_B))B^\top \;\approx\; B\Psi(Q_B)B^\top$. Indeed our numerical experiments show that Algorithm PN is particularly useful in situations where Q is "problematic", e.g. an empirical distribution of a sample with strong outliers.

4.3 Explicit pseudo-code

Standard M-estimators. Suppose that $Q = \sum_{i=1}^n w_i\delta_{x_i}$ with a certain weight vector $\boldsymbol{w} = (w_i)_{i=1}^n$ in $(0,1)^n$ such that $\sum_{i=1}^n w_i = 1$ and a data matrix $\boldsymbol{X} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^{n\times q}$. Then our Algorithm PN may be formulated as in Table 1.

Symmetrized M-estimators. Suppose that

$$Q \;=\; \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \delta_{x_i - x_j}$$

for a certain data matrix $\boldsymbol{X} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^{n\times q}$. In principle one could utilize the algorithm just described with $N = \binom{n}{2}$ in place of n and $\boldsymbol{X}$ replaced by a data matrix $\tilde{\boldsymbol{X}}$ containing all N differences $x_i - x_j$. For large n, however, this may require too much computer memory, and one should avoid the explicit storage of such a large data matrix $\tilde{\boldsymbol{X}}$.

$\Sigma \leftarrow \mathbf{AlgorithmPN}(\boldsymbol{X}, \boldsymbol{w}, \delta)$

$B \leftarrow \left(\sum_{i=1}^n w_i x_i x_i^\top\right)^{1/2}$
$\boldsymbol{Y} \leftarrow \boldsymbol{X} B^{-1}$
$\Psi \leftarrow \sum_{i=1}^n w_i \rho'(\|y_i\|^2)\, y_i y_i^\top$
$(U, \phi) \leftarrow \mathrm{Eigen}(\Psi)$

while $\|1_q - \phi\| > \delta$ **do**
    $B \leftarrow BU$
    $\boldsymbol{Y} \leftarrow \boldsymbol{Y} U$
    $\tilde{H} \leftarrow \mathrm{diag}(\phi) + \sum_{i=1}^n w_i \rho''(\|y_i\|^2) s(y_i) s(y_i)^\top$ $(+\, c\, 1_q 1_q^\top$ in Setting 0)
    $a \leftarrow \tilde{H}^{-1}(\phi - 1_q)$
    $\boldsymbol{Z} \leftarrow \boldsymbol{Y} \exp(-\mathrm{diag}(a)/2)$
    $DL \leftarrow \sum_{i=1}^n w_i \big(\rho(\|z_i\|^2) - \rho(\|y_i\|^2)\big) + \sum_{j=1}^q a_j$
    $DL_0 \leftarrow a^\top (1_q - \phi)/4$
    if $DL \le DL_0$ **then**
        $B \leftarrow B \exp(\mathrm{diag}(a)/2)$
        $\boldsymbol{Y} \leftarrow \boldsymbol{Z}$
    else
        $B \leftarrow B\, \mathrm{diag}(\phi)^{1/2}$
        $\boldsymbol{Y} \leftarrow \boldsymbol{Y}\, \mathrm{diag}(\phi)^{-1/2}$
    end if
    $\Psi \leftarrow \sum_{i=1}^n w_i \rho'(\|y_i\|^2)\, y_i y_i^\top$
    $(U, \phi) \leftarrow \mathrm{Eigen}(\Psi)$
end while

$\Sigma \leftarrow BB^\top$
return Σ

Table 1: Pseudo-code for the M-estimator.

It turned out that the computation time can be reduced substantially if we first compute the M-estimator $\mathbf{\Sigma}(\tilde{Q})$ for the surrogate distribution

$$\tilde{Q} \;:=\; \frac{1}{n}\sum_{i=1}^{n} \delta_{x_{\pi(i)}-x_{\pi(i+1)}}$$

with a randomly chosen permutation π of $\{1, 2, \ldots, n\}$ and $\pi(n+1) := \pi(1)$. Then we use this estimator $\mathbf{\Sigma}(\tilde{Q})$ as a starting parameter Σ_0 in Algorithm PN.

Table 2 contains pseudo-code for the computation of the symmetrized M-estimator without using a large data matrix $\tilde{\boldsymbol{X}}$. Instead it utilizes auxiliary programs to compute the following objects:

$$\begin{aligned}
\mathrm{RPermute}(n) \;&\to\; \text{a random permutation of } \{1, 2, \ldots, n\},\\
\mathrm{Psi}(\boldsymbol{X}) \;&\to\; \frac{1}{N}\sum_{1\le i<j\le n} \rho'(\|x_i - x_j\|^2)(x_i - x_j)(x_i - x_j)^\top,\\
\mathrm{H}(\phi, \boldsymbol{X}) \;&\to\; \mathrm{diag}(\phi) + \frac{1}{N}\sum_{1\le i<j\le n} \rho''(\|x_i - x_j\|^2)s(x_i - x_j)s(x_i - x_j)^\top,\\
\mathrm{DL}(\boldsymbol{X}, \boldsymbol{Y}, a) \;&\to\; \frac{1}{N}\sum_{1\le i<j\le n} \big[\rho(\|y_i - y_j\|^2) - \rho(\|x_i - x_j\|^2)\big] + \sum_{k=1}^{q} a_k.
\end{aligned}$$

5 Numerical examples and comparisons

In most of our simulation experiments we simulated data matrices $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^\top$ with independent rows $X_i = (X_{ij})_{j=1}^q$ having either standard Gaussian or standard Cauchy distribution on $\mathbb{R}^q$. In the latter case, $(X_{ij})_{j=1}^q$ is distributed as $(Z_j/Z_0)_{j=1}^q$ with independent random variables $Z_0, Z_1, \ldots, Z_q \sim \mathcal{N}(0, 1)$. In all experiments, iterations were stopped when the gradient $G_k = G(Q_k)$ of our target function satisfies $\|G_k\| \le 10^{-7}$, and the number of Monte Carlo simulations for each setting was 500.

The first three experiments were run on a MacBook Pro (2GHz Intel(R) Core i7, 16GB), the fourth experiment on a Windows server (two Intel(R) Xeon(R) CPU R5 2440 with 2.40GHz and 64GB). We used R 3.1.2 [14].

Comparisons in scatter-only settings. To compare the three algorithms FP, G and PN, we first implemented them in pure R code. Table 3 contains the mean number of iterations and the mean computing times for the scatter estimator $\mathbf{\Sigma}(\widehat{P})$ with $\rho = \rho_{1,q}$ based on a data matrix $\boldsymbol{X} \in \mathbb{R}^{500\times q}$,

$\Sigma \leftarrow \mathbf{AlgorithmPN.symm}(\boldsymbol{X}, \delta)$

$\pi \leftarrow \mathrm{RPermute}(n)$
$\boldsymbol{X}^0 \leftarrow [x_{\pi(1)} - x_{\pi(2)}, x_{\pi(2)} - x_{\pi(3)}, \ldots, x_{\pi(n)} - x_{\pi(1)}]^\top$
$B \leftarrow \mathbf{AlgorithmPN}(\boldsymbol{X}^0, (1/n)_{i=1}^n, \delta)^{1/2}$
$\boldsymbol{Y} \leftarrow \boldsymbol{X} B^{-1}$
$\Psi \leftarrow \mathrm{Psi}(\boldsymbol{Y})$
$(U, \phi) \leftarrow \mathrm{Eigen}(\Psi)$

while $\|1_q - \phi\| > \delta$ **do**
    $B \leftarrow BU$
    $\boldsymbol{Y} \leftarrow \boldsymbol{Y}U$
    $\tilde{H} \leftarrow \mathrm{H}(\phi, \boldsymbol{Y})$ $(+\, c\, 1_q 1_q^\top$ in Setting 0$)$
    $a \leftarrow \tilde{H}^{-1}(\phi - 1_q)$
    $\boldsymbol{Z} \leftarrow \boldsymbol{Y} \exp(-\operatorname{diag}(a)/2)$
    $DL \leftarrow \mathrm{DL}(\boldsymbol{Y}, \boldsymbol{Z}, a)$
    $DL_0 \leftarrow a^\top (1_q - \phi)/4$
    if $DL \leq DL_0$ **then**
        $B \leftarrow B \exp(\operatorname{diag}(a)/2)$
        $\boldsymbol{Y} \leftarrow \boldsymbol{Z}$
    else
        $B \leftarrow B \operatorname{diag}(\phi)^{1/2}$
        $\boldsymbol{Y} \leftarrow \boldsymbol{Y} \operatorname{diag}(\phi)^{-1/2}$
    end if
    $\Psi \leftarrow \mathrm{Psi}(\boldsymbol{Y})$
    $(U, \phi) \leftarrow \mathrm{Eigen}(\Psi)$
end while

$\Sigma \leftarrow BB^\top$
return Σ

Table 2: Pseudo-code for the symmetrized M-estimator.

	Gaussian data			Cauchy data		
Algorithm	FP	G	PN	FP	G	PN
$\boldsymbol{q = 5}$						
Iterations	83.9 (2)	31.2 (4)	5.1 (0)	116.4 (3)	45.5 (14)	8.5 (1)
Time [ms]	13.5 (0.5)	11.4 (1.8)	1.8 (0.3)	18.5 (1.0)	16.8 (5.3)	2.8 (0.5)
Relative efficiency	FP	1.18 G	**7.71** **6.51**	FP	1.10 G	**6.53** **5.95**
$\boldsymbol{q = 10}$						
Iterations	141.6 (1)	46.0 (6)	6.0 (0)	189.4 (3)	69.4 (30)	9.3 (1)
Time [ms]	41.9 (1.0)	25.0 (2.8)	3.1 (0.3)	56.2 (2.3)	37.1 (16.0)	5.0 (1.0)
Relative efficiency	FP	1.68 G	**13.37** **7.97**	FP	1.51 G	**11.19** **7.40**
$\boldsymbol{q = 20}$						
Iterations	252.2 (2)	119.2 (6)	6.0 (0)	332.2 (4)	103.7 (43)	10.6 (1)
Time [ms]	176.2 (4.8)	120.2 (7.8)	6.9 (0.3)	230.1 (4.8)	104.4 (43.4)	12.4 (1.3)
Relative efficiency	FP	1.47 G	**25.65** **17.49**	FP	2.20 G	**18.54** **8.41**

Table 3: Computation costs and relative efficiencies in scatter-only settings ($n = 500$, $\nu = 1$).

$q = 5, 10, 20$. The table entries are the mean iteration numbers and mean computations times in milliseconds [ms]. In brackets the corresponding inter quartile ranges are recorded as well. The relative efficiencies are the ratios of the mean computation times. Algorithm G is already more efficient than Algorithm FP, but obviously Algorithm PN is substantially faster than the other two, and this advantage grows with the dimension q. Note also that computation costs are higher for Cauchy data than for Gaussian data.

Comparisons in location-scatter settings. Now we consider the empirical distribution $\widehat{P}$ of the rows of $\boldsymbol{X}$ and for given $\nu \geq 1$ the minimizer $\big(\boldsymbol{\mu}_\nu(\widehat{P}), \boldsymbol{\Sigma}_\nu(\widehat{P})\big)$ of

$$L_\nu(\mu, \Sigma, \widehat{P}) \;:=\; L_\nu(\Gamma(\mu, \Sigma), \widehat{Q})$$

over all $(\mu, \Sigma) \in \mathbb{R}^q \times \mathbb{R}^{q\times q}_{\mathrm{sym},>0}$. Here $\Gamma(\mu, \Sigma) \in \mathbb{R}^{(q+1)\times(q+1)}_{\mathrm{sym},>0}$ is defined as in (2), $\widehat{Q}$ stands for the empirical distribution of the augmented data points $[X_i^\top, 1]^\top \in \mathbb{R}^{q+1}$, $1 \leq i \leq n$, and

$$L_\nu(\Gamma, \widehat{Q}) \;:=\; \int \big[\rho_{\nu-1,q+1}(y^\top \Gamma^{-1} y) - \rho_{\nu-1,q+1}(y^\top y)\big]\, \widehat{Q}(dy) + \log\det(\Gamma)$$

for arbitrary $\Gamma \in \mathbb{R}^{(q+1)\times(q+1)}_{\mathrm{sym},>0}$.

In principle, we may apply any of the three algorithms FP, G and PN to the empirical distribution $\widehat{Q}$ to compute a minimizer $\widehat{\Gamma}$ of $L_\nu(\cdot, \widehat{Q})$. In case of $\nu > 1$ this minimizer satisfies automatically $\widehat{\Gamma}_{q+1,q+1} = 1$, so $\widehat{\Gamma} = \Gamma\big(\boldsymbol{\mu}_\nu(\widehat{P}), \boldsymbol{\Sigma}_\nu(\widehat{P})\big)$. In case of $\nu = 1$, $\widehat{\Gamma}$ equals $\Gamma\big(\boldsymbol{\mu}_\nu(\widehat{P}), \boldsymbol{\Sigma}_\nu(\widehat{P})\big)$ times $\widehat{\Gamma}_{q+1,q+1}$.

In addition we implemented a variant FP_3 of FP proposed by Arslan et al. [1]. Suppose that $(\mu_k, B_k B_k^\top)$ with nonsingular $B_k \in \mathbb{R}^{q\times q}$ is a current candidate for $\big(\boldsymbol{\mu}_\nu(\widehat{P}), \boldsymbol{\Sigma}_\nu(\widehat{P})\big)$. Let $\widehat{Q}_k$ denote the empirical distribution of the standardized data points $B_k^{-1}(X_i - \mu_k)$, $1 \le i \le n$, augmented by an additional component 1, and define

$$\Psi_k \;:=\; \int \rho'_{\nu-1,q+1}(y^\top y) y y^\top \, \widehat{Q}_k(dy).$$

Recall that $(\mu_k, B_k B_k^\top)$ equals $\big(\boldsymbol{\mu}_\nu(\widehat{P}), \boldsymbol{\Sigma}_\nu(\widehat{P})\big)$ if, and only if, $\Psi_k = I_{q+1}$. Now we write $\Psi_k = \lambda_k \Gamma(\delta_k, C_k C_k^\top)$ for some $\lambda_k > 0$, $\delta_k \in \mathbb{R}^k$ and a nonsingular matrix $C_k \in \mathbb{R}^{q\times q}$. Then the next candidate for $\big(\boldsymbol{\mu}_\nu(\widehat{P}), \boldsymbol{\Sigma}_\nu(\widehat{P})\big)$ equals $(\mu_{k+1}, B_{k+1} B_{k+1}^\top)$ with

$$\mu_{k+1} \;:=\; \mu_k + B_k \delta_k, \quad B_{k+1} \;:=\; B_k C_k.$$

To provide a fair comparison, we used the same stopping criterion as for the other algorithms, that means, we considered the norm of $I_{q+1} - \Psi_k$.

For $n = 100$ and $q = 10$ we simulated data matrices $\boldsymbol{X} \in \mathbb{R}^{n\times q}$ with independent entries

$$X_{ij} \;\sim\; \begin{cases} \mathcal{N}(\delta, 1) & \text{if } i \le n/10 \text{ and } j = 1, \\ \mathcal{N}(0, 1) & \text{else,} \end{cases}$$

where $\delta \ge 0$ is a certain parameter quantifying the outlyingness of the $n/10$ first data vectors. The left and right half of Table 4 show the resulting computation costs and times for $\delta = 0, 10, 20$ when $\nu = 1$ and $\nu = 2$, respectively. For $\nu = 1$, algorithm FP is more efficient than FP_3. Indeed one can easily verify that the two algorithms are essentially equivalent, the only difference being how they factorize matrices such as Ψ_k. For $\delta = 0$, algorithm FP ($\nu = 1$) and algorithm FP_3 ($\nu = 2$) are remarkably efficient and even outperform algorithm PN. But for larger values of δ, leading to heterogeneous data sets, PN is clearly the fastest method.

Comparisons for symmetrized scatter estimators, I. As mentioned in the introduction, computation time becomes a major issue when computing symmetrized scatter estimators. In the simulation experiments described below we simulated data matrices $\boldsymbol{X} \in \mathbb{R}^{n\times q}$ with independent rows following a multivariate standard Gaussian or standard Cauchy distribution on $\mathbb{R}^q$.

	$\nu = 1$			$\nu = 2$		
Algorithm	FP	FP_3	PN	FP	FP_3	PN
$\boldsymbol{\delta = 0}$						
Iterations	15.1 (0)	15.1 (0)	9.6 (1)	152.0 (3)	13.8 (1)	8.9 (0)
Time [ms]	2.3 (0.2)	2.7 (0.2)	3.0 (0.2)	21.8 (0.6)	2.8 (0.3)	2.9 (0.1)
Relative efficiency	FP	0.87 FP_3	**0.77** **0.88**	FP	7.81 FP_3	**7.62** **0.98**
$\boldsymbol{\delta = 10}$						
Iterations	27.4 (4)	27.4 (4)	12.3 (1)	157.3 (3)	25.8 (3)	11.6 (1)
Time [ms]	4.0 (0.6)	4.7 (0.7)	3.7 (0.3)	22.3 (0.6)	4.9 (0.6)	3.7 (0.3)
Relative efficiency	FP	0.85 FP_3	**1.09** **1.28**	FP	4.60 FP_3	**6.11** **1.33**
$\boldsymbol{\delta = 20}$						
Iterations	47.2 (6)	47.2 (6)	17.2 (2)	161.4 (3)	42.0 (4)	15.6 (1)
Time [ms]	6.6 (0.9)	7.8 (1.0)	5.0 (0.5)	23.0 (0.6)	7.9 (1.0)	4.9 (0.4)
Relative efficiency	FP	0.84 FP_3	**1.31** **1.56**	FP	2.93 FP_3	**4.66** **1.59**

Table 4: Computation costs and relative efficiencies in location-scatter settings ($q = 10$, $n = 100$).

Our first simulation experiment concerns 2×2 different variants of Algorithm PN for symmetrized estimators with $\rho = \rho_{q,1}$: On the one hand we compared storing all $N = n(n-1)/2$ pairwise differences of data vectors in a big matrix and running the algorithm in Table 1 ("PN-all") with a less memory-intensive version where all statistics are computed sequentially as in Table 2 ("PN-seq"). In both cases we first prewhitened the data by means of a scatter estimator based on n randomly chosen pairs of observations, see the first four lines of pseudo-code in Table 2. On the other hand we investigated the benefits of the latter prewhitening step and implemented versions without it ("PN-all.0" and "PN-seq.0"). Figures 1 and 2 show box plots of the computation times (using pure R code) for dimension $q = 10$ and sample sizes $n = 100$ and $n = 500$, respectively. One sees clearly that for small to moderate sample sizes version "PN-all" is faster than "PN-seq". But for larger sample sizes "PN-seq" becomes clearly preferable. Comparing "PN-all.0" with "PN-all" and "PN-seq.0" with "PN-seq" shows that prewhitening is particularly beneficial for the heavy-tailed distribution and larger sample sizes. Note that all computation times for the symmetrized scatter estimators are in seconds [s] rather than milliseconds [ms] as before.

More efficient code. The new algorithms described in this paper are implemented in the R package *fastM* (Dümbgen et al. [4]) which is publicly available on CRAN. This includes implemen-

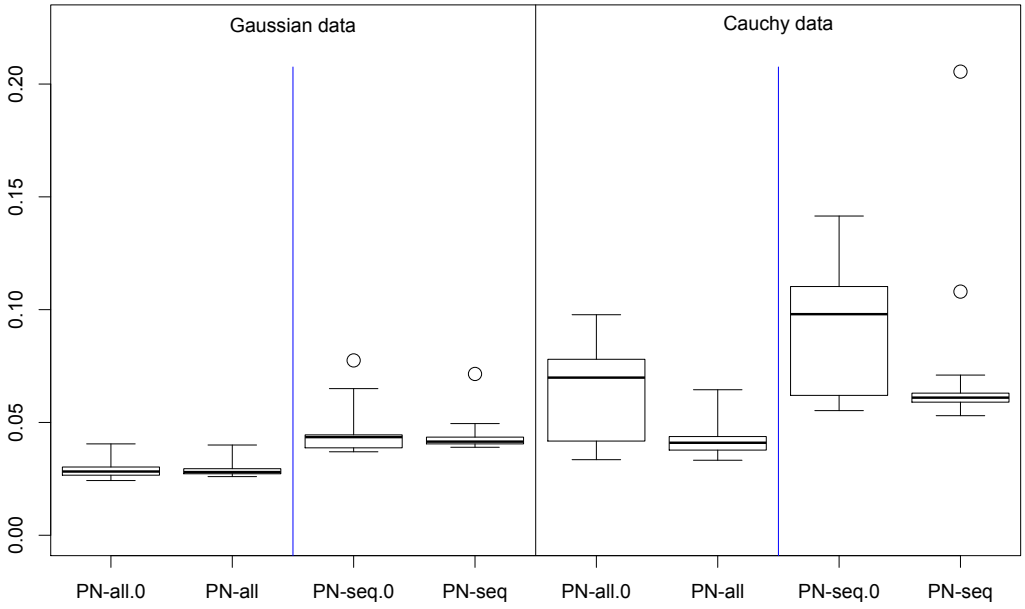

Figure 1: Computation times [s] of four variants of AlgorithmPN.symm ($q = 10, n = 100, \nu = 1$).

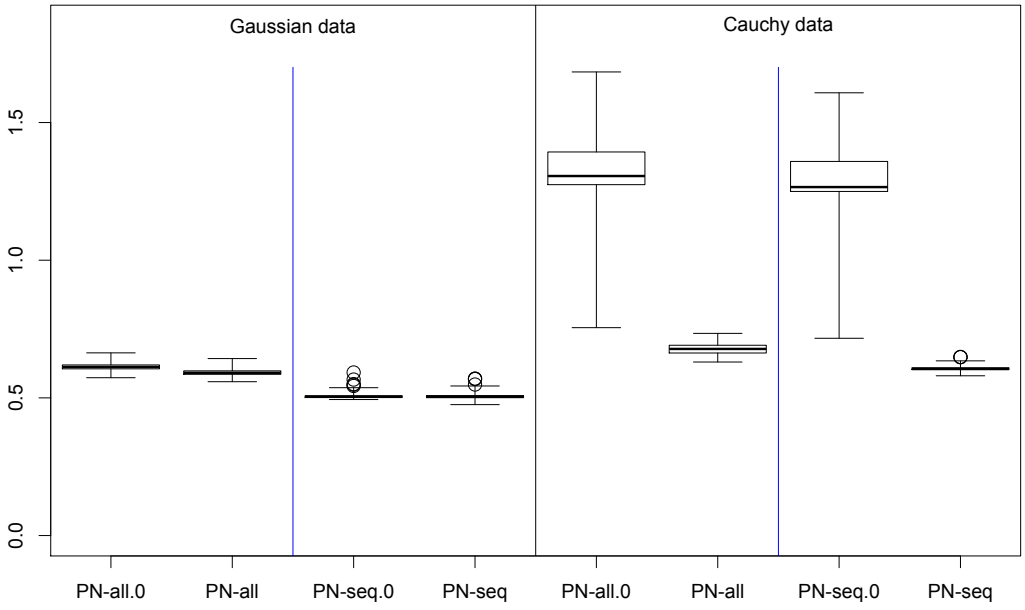

Figure 2: Computation times [s] of four variants of AlgorithmPN.symm ($q = 10, n = 500, \nu = 1$).

	Gaussian data				Cauchy data			
	Iter.	Time [s] R	Time [s] C++	Rel. eff.	Iter.	Time [s] R	Time [s] C++	Rel. eff.
$\boldsymbol{\nu = 0}$								
$q = 5$	4.0 (0)	1.2 (0.3)	0.2 (0.1)	**6.81**	5.1 (0)	1.3 (0.2)	0.2 (0.1)	**5.67**
$q = 10$	5.0 (0)	1.7 (0.4)	0.4 (0.2)	**3.91**	6.0 (0)	2.1 (0.4)	0.5 (0.3)	**4.04**
$q = 20$	5.0 (0)	2.9 (0.7)	0.9 (0.3)	**3.13**	6.9 (0)	3.7 (1.0)	1.2 (0.3)	**3.15**
$\boldsymbol{\nu = 1}$								
$q = 5$	4.0 (0)	1.2 (0.3)	0.2 (0.1)	**6.40**	5.1 (0)	1.3 (0.2)	0.2 (0.2)	**5.44**
$q = 10$	5.0 (0)	1.7 (0.4)	0.4 (0.2)	**3.96**	6.0 (0)	2.0 (0.4)	0.5 (0.3)	**3.97**
$q = 20$	5.0 (0)	2.9 (0.8)	0.9 (0.3)	**3.11**	6.9 (0)	3.7 (1.0)	1.2 (0.4)	**3.11**

Table 5: Computation costs and relative efficiencies for symmetrized scatter ($n = 500$).

tations with C++ code which are even more efficient. We did substantial simulation experiments to compare our package with other implementations of M-estimators, namely (i) the function *cov.trob* in the package *MASS* (Venables and Ripley [19]) and (ii) the function *tM* in the package *ICS* (Nordhausen et al. [11]). Both functions are essentially fix-point approaches. In particular, *tM* is based on a maximum-likelihood and EM interpretation of the fixed point equation and uses algorithm FP_3 by Arslan et al. [1] mentioned before. All in all our new algorithms were always comparable, often faster and in some settings even substantially faster than the other methods. A fair comparison is difficult, though, because the established algorithms use different stopping criteria. Both *cov.trob* and *tM* update the location and scatter parameters separately and do not treat it as our algorithms do, as a scatter-only problem. For the symmetrized estimator with $\rho = \rho_{0,q}$, there is the function duembgen.shape available in the R package *ICSNP* (Nordhausen et al. [12]), which is essentially Algorithm FP and utilizes R and C code.

Comparisons for symmetrized scatter estimators, II. Finally, Tables 5 and 6 compare the performance of the symmetrized estimator as implemented in *fastM* with pure R code and with C++ code, where $\rho = \rho_{\nu,q}$, $\nu = 0, 1$. The results show that Algorithm PN with C++ code is substantially faster than its pure R version.

	Gaussian data				Cauchy data			
	Iter.	Time [s] R	Time [s] C++	Rel. eff.	Iter.	Time [s] R	Time [s] C++	Rel. eff.
$\boldsymbol{\nu = 0}$								
$q = 5$	3.2 (0)	7.9 (1.6)	1.9 (0.5)	**4.03**	4.0 (0)	9.5 (1.4)	2.3 (0.5)	**4.06**
$q = 10$	4.0 (0)	14.3 (2.6)	4.3 (0.5)	**3.30**	4.6 (1)	16.0 (3.5)	4.9 (1.0)	**3.27**
$q = 20$	4.0 (0)	33.1 (7.9)	10.1 (0.2)	**3.28**	5.0 (0)	40.7 (8.3)	12.2 (0.2)	**3.33**
$\boldsymbol{\nu = 1}$								
$q = 5$	3.2 (0)	7.7 (1.4)	1.9 (0.4)	**3.99**	4.0 (0)	9.5 (1.4)	2.4 (0.5)	**4.00**
$q = 10$	4.0 (0)	14.3 (2.8)	4.4 (0.6)	**3.24**	4.7 (1)	16.2 (3.4)	5.0 (0.5)	**3.25**
$q = 20$	4.0 (0)	33.1 (7.7)	10.1 (0.2)	**3.27**	5.0 (0)	40.8 (7.9)	12.3 (0.2)	**3.32**

Table 6: Computation costs and relative efficiencies for symmetrized scatter ($n = 2000$).

6 Proofs

Proof of Corollaries 6 and 7. For $t \in \mathbb{R}$ define $F(t) := L(B\exp(tA)B^\top, Q)$ and $B(t) := B\exp((t/2)A)$. Note that $B(t)$ is nonsingular with $B(0) = B$. For $u \in \mathbb{R}$,

$$\begin{aligned} F(t+u) - F(t) &= L(B(t)\exp(uA)B(t)^\top, Q) - L(B(t)B(t)^\top, Q) \\ &= L(\exp(uA), Q_{B(t)}) \\ &= uG(A, Q_{B(t)}) + \frac{u^2}{2}H(A, Q_{B(t)}) + o(u^2) \end{aligned}$$

as $u \to 0$. Since both $G(A, Q_{B(t)})$ and $H(A, Q_{B(t)})$ are continuous in $t \in \mathbb{R}$, this expansion shows that F is twice continuously differentiable with $F'(t) = G(A, Q_{B(t)})$ and

$$F''(t) = H(A, Q_{B(t)}) \begin{cases} \geq 0, & \\ > 0 & \text{in Setting 0 if } A \neq 0, \operatorname{tr}(A) = 0, \\ > 0 & \text{in Setting 1 if } A \neq 0. \end{cases}$$

In particular, F is convex. It is even strictly convex unless

$$\begin{cases} A = sI_q \text{ for some } s \in \mathbb{R} & \text{in Setting 0,} \\ A = 0 & \text{in Setting 1.} \end{cases}$$

To verify Corollary 7, we utilize the same auxiliary function $F = F(\cdot \mid B, A)$ and write $L(B\exp(A)B^\top, Q) - L(BB^\top, Q)$ as

$$F(1) - F(0) = F'(0) + \int_0^1 (1-t)F''(t)\,dt = G(A, Q_B) + \int_0^1 (1-t)H(A, Q_{B(t)})\,dt.$$

Now let $B = \Sigma^{1/2}V$ with an orthogonal matrix $V \in \mathbb{R}^{q\times q}$, and define

$$C(t) \;:=\; B(t)V^\top \;=\; \Sigma^{1/2}V\exp((t/2)A)V^\top.$$

Then

$$\begin{aligned}
r(B,A) \;&=\; \|A\|^{-2}\int_0^1 (1-t)\big(H(A,Q_{B(t)}) - H(A,Q_B)\big)\,dt\\
&=\; \|A\|^{-2}\int_0^1 (1-t)\big(H(V^\top AV,Q_{C(t)}) - H(V^\top AV,Q_{\Sigma^{1/2}})\big)\,dt,\\
r_*(B,A) \;&=\; \|A\|^{-2}\int_0^1 (1-t)\big(H(A,Q_{B(t)}) - H(V^\top AV,Q_*)\big)\,dt\\
&=\; \|A\|^{-2}\int_0^1 (1-t)\big(H(V^\top AV,Q_{C(t)}) - H(V^\top AV,Q_*)\big)\,dt,
\end{aligned}$$

so $|r(B,A)| + |r_*(B,A)|$ is no larger than $3/2$ times the supremum of

$$\big|H(A',Q_{\Sigma^{1/2}V_o\exp(A_o)V_o^\top}) - H(A',Q_*)\big|$$

over all $A', A_o \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$ with $\|A'\| \le 1$, $\|A_o\| \le \|A\|/2$ and all orthogonal matrices $V_o \in \mathbb{R}^{q\times q}$. But this converges to zero as $\Sigma = BB^\top \to \mathbf{\Sigma}(Q)$ and $A \to 0$, because then

$$\begin{aligned}
\big\|\Sigma^{1/2}V_o\exp(A_o)V_o^\top - \mathbf{\Sigma}(Q)^{1/2}\big\| \;&\le\; \|\Sigma^{1/2}\|\|V_o\exp(A_o)V_o^\top - I_q\| + \|\Sigma^{1/2} - \mathbf{\Sigma}(Q)^{1/2}\|\\
&=\; \|\Sigma^{1/2}\|\|\exp(A_o) - I_q\| + \|\Sigma^{1/2} - \mathbf{\Sigma}(Q)^{1/2}\|\\
&\to\; 0.
\end{aligned}$$

Finally, because $G(Q_*) = I_q - \Psi(Q_*) = 0$, we may write

$$\begin{aligned}
H(V^\top AV,Q_*) \;&=\; \big\langle (V^\top AV)^2, I_q\big\rangle + \int \rho''(\|x\|^2)(x^\top V^\top AVx)^2\,Q_*(dx)\\
&=\; \|A\|^2 + \int \rho''(\|x\|^2)(x^\top V^\top AVx)^2\,Q_*(dx).
\end{aligned}$$

□

Proof of Theorem 8. Dropping the index k for the moment, suppose that $\Sigma = BB^\top$ is our current candidate parameter. Then one step of Algorithm FP replaces Σ with

$$B\Psi(Q_B)B^\top \;=\; \int \rho'(x^\top\Sigma^{-1}x)xx^\top\,Q(dx).$$

Hence $L(\Sigma,Q)$ changes by

$$\delta_1(\Sigma) \;:=\; L(B\Psi(Q_B)B^\top,Q) - L(\Sigma,Q) \;=\; L(\Psi(Q_B),Q_B) \;\le\; 0,$$

and the inequality is strict unless Σ minimizes $L(\cdot, Q)$ already, see (6). Note also that $\delta_1(\Sigma)$ is a continuous function of Σ.

Algorithm PN is slightly more difficult to quantify, because the eigenmatrix U in the representation $\Psi(Q_B) = U\,\mathrm{diag}(\phi)U^\top$ is not unique. However,

$$\begin{aligned}\min_{a\in\mathbb{R}^q}\Big(\tilde G(Q_{BU})^\top a + \frac{1}{2}a^\top \tilde H(Q_{BU})a\Big) \;&\le\; \min_{a\in\mathrm{span}(\tilde G(Q_{BU}))}\Big(\tilde G(Q_{BU})^\top a + \frac{1}{2}a^\top \tilde H(Q_{BU})a\Big)\\ &=\; \frac{-\|\tilde G(Q_{BU})\|^2}{2\tilde G(Q_{BU})^\top \tilde H(Q_{BU})\tilde G(Q_{BU})}\\ &=\; \frac{-\|G(Q_{BU})\|^2}{2H(G(Q_{BU}),Q_{BU})}\\ &=\; \frac{-\|G(Q_{\Sigma^{1/2}})\|^2}{2H(G(Q_{\Sigma^{1/2}}),Q_{\Sigma^{1/2}})}.\end{aligned}$$

In the last step we utilized that fact that $BU = \Sigma^{1/2}W$ for some orthogonal matrix $W \in \mathbb{R}^{q\times q}$, and that $G(Q_{BU}) = W^\top G(Q_{\Sigma^{1/2}})W$, $H(G(Q_{BU}),Q_{BU}) = H(G(Q_{\Sigma^{1/2}}),Q_{\Sigma^{1/2}})$. Consequently, the change of $L(\Sigma,Q)$ with Algorithm PN is at least

$$\delta_2(\Sigma) \;:=\; \max\Big(\delta_1(\Sigma), \frac{-\|G(Q_{\Sigma^{1/2}})\|^2}{4H(G(Q_{\Sigma^{1/2}}),Q_{\Sigma^{1/2}})}\Big) \;\le\; 0,$$

again a continuous function of Σ, and the inequality is strict unless Σ minimizes $L(\cdot,Q)$.

In Setting 1, the minimizer $\mathbf{\Sigma}_\rho(Q)$ is unique, and we may utilize the following standard arguments: Suppose that $(\Sigma_k)_k$ does not converge to $\mathbf{\Sigma}_\rho(Q)$. We know that $L(\Sigma_k,Q)$ is decreasing in $k\ge 0$, and all Σ_k belong to the compact set $\{\Sigma : L(\Sigma,Q)\le L(\Sigma_0,Q)\}$. Hence there would exist a subsequence $(\Sigma_{k(\ell)})_\ell$ with limit $\Sigma_* \ne \mathbf{\Sigma}_\rho(Q)$. But then continuity of $L(\cdot,Q)$ and $\delta_j(\cdot)$ would imply that

$$\begin{aligned}L(\Sigma_*,Q) \;&=\; \lim_{\ell\to\infty} L(\Sigma_{k(\ell)},Q)\\ &=\; \lim_{\ell\to\infty} L(\Sigma_{k(\ell)+1},Q)\\ &\le\; \lim_{\ell\to\infty}\big(L(\Sigma_{k(\ell)},Q)+\delta_j(\Sigma_{k(\ell)})\big)\\ &=\; L(\Sigma_*,Q)+\delta_j(\Sigma_*)\\ &<\; L(\Sigma_*,Q).\end{aligned}$$

In Setting 0, note first that $L(\Sigma,Q)$, $\Psi(Q_B)$ and $H(Q_B)$ remain unchanged if we replace (Σ,B) with $(t\Sigma,t^{1/2}B)$ for some number $t>0$. Hence, with the same arguments as in Setting 1, we may conclude that $t_k\Sigma_k \to \mathbf{\Sigma}_0(Q)$ as $k\to\infty$, where $t_k := \det(\Sigma_k)^{-q/2}$.

Now in case of Algorithm FP an elementary calculation shows that the matrices $M_k := \mathbf{\Sigma}_0(Q)^{-1/2}\Sigma_k\mathbf{\Sigma}_0(Q)^{-1/2}$ satisfy the equation

$$M_{k+1} \;=\; \int \frac{q}{x^\top M_k^{-1} x}\, xx^\top \, Q_{\mathbf{\Sigma}_0(Q)^{1/2}}(dx).$$

Together with the equation $\Psi(Q_{\mathbf{\Sigma}_0(Q)^{1/2}}) = I_q$ this implies that

$$\lambda_{\min}(M_{k+1}) \;\geq\; \lambda_{\min}(M_k) \quad \text{and} \quad \lambda_{\max}(M_{k+1}) \;\leq\; \lambda_{\max}(M_k).$$

Hence the sequence $(M_k)_k$ converges to a multiple of the identity matrix. In other words, $(\Sigma_k)_k$ converges to a multiple of $\mathbf{\Sigma}_0(Q)$.

The definition of Algorithm PN implies that for sufficiently large k, the new candidate Σ_{k+1} is given by $B_k \exp(\mathrm{diag}(a_k))B_k^\top$ with $a_k \in \mathbb{R}^q$ satisfying $1_q^\top a_k = 0$. Hence $\det(\Sigma_{k+1}) = \det(\Sigma_k)$ for sufficiently large k. Consequently $(\Sigma_k)_k$ converges to a multiple of $\mathbf{\Sigma}_0(Q)$. □

Proof of Lemma 9. The fact that $\mathbf{\Sigma}(Q)$ is a positive multiple of Σ_o follows from simple equivariance considerations as outlined in [5]. Now let $\mathbf{\Sigma}(Q) = CC^\top$ with nonsingular $C \in \mathbb{R}^{q\times q}$, and let $Z := C^{-1}X$ with $X \sim Q$. The random vector Z has a spherically symmetric distribution around 0 in the sense that for any orthogonal matrix $V \in \mathbb{R}^{q\times q}$, the distributions of $V^\top Z$ and Z coincide. We may write

$$\begin{aligned}\Psi(Q_B) \;&=\; \mathbb{E}\big[\rho'(\|B^{-1}X\|^2)(B^{-1}X)(B^{-1}X)^\top\big] \\ &=\; B^{-1}C\,\mathbb{E}\big[\rho'(Z^\top C^\top\Sigma^{-1}CZ)ZZ^\top\big]C^\top B^{-\top}.\end{aligned}$$

Next let

$$C^\top\Sigma^{-1}C \;=\; V\,\mathrm{diag}(\gamma)V^\top$$

with an orthogonal matrix $V \in \mathbb{R}^{q\times q}$ and a vector $\gamma \in (0,\infty)^q$ containing the eigenvalues of $C^\top\Sigma^{-1}C$, i.e. the eigenvalues of $\Sigma^{-1}\mathbf{\Sigma}(Q)$. Then

$$B^{-1}C \;=\; \tilde{U}\,\mathrm{diag}(\gamma)^{1/2}V^\top$$

for another orthogonal matrix $\tilde{U}$, so

$$\begin{aligned}
\Psi(Q_B) &= \tilde{U}\,\mathrm{diag}(\gamma)^{1/2}V^\top\,\mathbb{E}\big[\rho'(Z^\top V\,\mathrm{diag}(\gamma)V^\top Z)ZZ^\top\big]V\,\mathrm{diag}(\gamma)^{1/2}\tilde{U}^\top \\
&= \tilde{U}\,\mathrm{diag}(\gamma)^{1/2}\,\mathbb{E}\big[\rho'((V^\top Z)^\top\,\mathrm{diag}(\gamma)(V^\top Z))(V^\top Z)(V^\top Z)^\top\big]\,\mathrm{diag}(\gamma)^{1/2}\tilde{U}^\top \\
&= \tilde{U}\,\mathrm{diag}(\gamma)^{1/2}\,\mathbb{E}\big[\rho'(Z^\top\,\mathrm{diag}(\gamma)Z)ZZ^\top\big]\,\mathrm{diag}(\gamma)^{1/2}\tilde{U}^\top \\
&= \tilde{U}\,\mathrm{diag}(\gamma)^{1/2}\,\mathbb{E}\Big[\rho'\Big(\sum_{i=1}^q\gamma_i Z_i^2\Big)(Z_jZ_k)_{j,k=1}^q\Big]\,\mathrm{diag}(\gamma)^{1/2}\tilde{U}^\top \\
&= \tilde{U}\,\mathrm{diag}(\gamma)^{1/2}\,\mathbb{E}\Big[\rho'\Big(\sum_{i=1}^q\gamma_i Z_i^2\Big)\,\mathrm{diag}\big((Z_j^2)_{j=1}^q\big)\Big]\,\mathrm{diag}(\gamma)^{1/2}\tilde{U}^\top \\
&= \tilde{U}\,\mathbb{E}\Big[\rho'\Big(\sum_{i=1}^q\gamma_i Z_i^2\Big)\,\mathrm{diag}\big((\gamma_jZ_j^2)_{j=1}^q\big)\Big]\tilde{U}^\top,
\end{aligned}$$

by spherical symmetry of the distribution of Z. Hence

$$\Psi(Q_B) \;=\; \tilde{U}\,\mathrm{diag}(\phi)\tilde{U}^\top$$

with $\phi \in (0,\infty)^q$ given by

$$\phi_j \;:=\; \mathbb{E}\Big(\rho'\Big(\sum_{i=1}^q\gamma_i Z_i^2\Big)\gamma_j Z_j^2\Big).$$

Moreover, since $\rho' > 0$ and the distribution of $(Z_i^2)_{i=1}^q$ is invariant under permuting the components of Z,

$$\phi_j = \phi_k \;\text{ if, and only if, }\; \gamma_j = \gamma_k.$$

One may also say that ϕ is the unique vector of eigenvalues of $\Psi(Q_B)$, and the columns $\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_q$ of $\tilde{U}$ are corresponding eigenvectors. If we consider another spectral decomposition $\Psi(Q_B) = U\,\mathrm{diag}(\phi)U^\top$ with U having orthonormal columns $u_1, u_2, \ldots, u_q$, then

$$U\exp(\mathrm{diag}(a))U^\top \;=\; \tilde{U}\exp(\mathrm{diag}(a))\tilde{U}^\top$$

for any vector $a \in \mathbb{R}^q$ such that $a_j = a_k$ whenever $\phi_j = \phi_k$. In particular, if we choose $a := \big(\log(\gamma_j)\big)_{j=1}^q$, then

$$\begin{aligned}
&BU\exp(\mathrm{diag}(a))U^\top B^\top \\
&\quad= B\tilde{U}\,\mathrm{diag}(\gamma)\tilde{U}^\top B^\top \;=\; B(B^{-1}C)(B^{-1}C)^\top B^\top \;=\; CC^\top \;=\; \boldsymbol{\Sigma}(Q).
\end{aligned}$$

□

Acknowledgement. The authors are grateful to Mathias Drton for his interest and questions which led to Lemma 9. We are also indebted to an anonymous referee for detailed and constructive comments.

References

[1] O. ARSLAN, J. KENT, AND P. CONSTABLE, *Convergence behaviour of the EM algorithm for the t-distribution*, Comm. Statist. A Theory Meth., 24 (1995), pp. 2981–3000.

[2] R. M. DUDLEY, S. SIDENKO, AND Z. WANG, *Differentiability of t-functionals of location and scatter*, Ann. Statist., 37 (2009), pp. 939–960.

[3] L. DÜMBGEN, *On Tyler's M-functional of scatter in high dimension*, Ann. Inst. Statist. Math., 50 (1998), pp. 471–491.

[4] L. DÜMBGEN, K. NORDHAUSEN, AND H. SCHUHMACHER, *fastM: Fast Computation of Multivariate M-estimators*, 2014. R package version 0.0-1.

[5] L. DÜMBGEN, M. PAULY, AND T. SCHWEIZER, *M-functionals of multivariate scatter*, Stat. Surv., 9 (2015), pp. 32–105.

[6] P. J. HUBER, *Robust covariances*, in Statistical decision theory and related topics, II (Proc. Sympos., Purdue Univ., Lafayette, Ind., 1976), Academic Press, New York, 1977, pp. 165–191.

[7] P. J. HUBER, *Robust Statistics*, Wiley, New York, 1981.

[8] J. T. KENT AND D. E. TYLER, *Redescending M-estimates of multivariate location and scatter*, Ann. Statist., 19 (1991), pp. 2102–2119.

[9] J. T. KENT, D. E. TYLER, AND Y. VARDI, *A curious likelihood identity for the multivariate t-distribution*, Comm. Statist. Sim. Comp., 23 (1994), pp. 441–453.

[10] K. NORDHAUSEN, H. OJA, AND E. OLLILA, *Robust independent component analysis based on two scatter matrices*, Austrian J. Statist., 37 (2008), pp. 91–100.

[11] K. NORDHAUSEN, H. OJA, AND D. E. TYLER, *Tools for exploring multivariate data: The package ICS*, Journal of Statistical Software, 28 (2008), pp. 1–31.

[12] K. NORDHAUSEN, S. SIRKIA, H. OJA, AND D. E. TYLER, *ICSNP: Tools for Multivariate Nonparametrics*, 2012. R package version 1.0-9.

[13] K. NORDHAUSEN AND D. E. TYLER, *A cautiounary note on robust covariance plug-in methods*, Biometrika, 102 (2015), pp. 573–588.

[14] R CORE TEAM, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[15] R. A. REDNER AND H. F. WALKER, *Mixture densities, maximum likelihood and the EM algorithm*, SIAM Review, 26 (1984), pp. 195–239.

[16] S. SIRKIÄ, S. TASKINEN, AND H. OJA, *Symmetrised M-estimators of multivariate scatter*, J. Multivar. Anal., 98 (2007), pp. 1611–1629.

[17] D. E. TYLER, *A distribution-free M-estimator of multivariate scatter*, Ann. Statist., 15 (1987), pp. 234–251.

[18] D. E. TYLER, F. CRITCHLEY, L. DÜMBGEN, AND H. OJA, *Invariant coordinate selection (with discussion)*, J. Royal Statist. Soc. B, 71 (2009), pp. 549–592.

[19] W. N. VENABLES AND B. D. RIPLEY, *Modern Applied Statistics with S*, Springer, New York, fourth ed., 2002. ISBN 0-387-95457-0.